From CTRAN to WordWave: More CTS for Less

Owen Kimball, Rukmini Iyer, John Makhoul

January CTRAN Experiment

- CTS training data can be transcribed faster (~10xRT) and at lower cost ($250/hr), plus post processing, with no noticeable effect on recognition accuracy
- Adding 80 hrs of data transcribed in this manner to existing 290 hrs of CTS training data resulted in lower word error rate
- Want to transcribe all of remaining Switchboard 2 and Fisher data
- But,
  - CTRAN cannot increase throughput easily
  - We would like to lower transcription costs further

WordWave

- **Transcription company headquartered in Boston, with main offices in UK and New Zealand**
  - **Can produce good-quality transcripts fast**
  - **They include capitalization and punctuation**
  - **High throughput promised (100+ hrs of speech per week)**
  - **Lower cost: ~$165/hr of speech**
- **Modified their Style Guide for our purposes**
  - **Went through several iterations**
  - **Very responsive to our requests**
  - **We now have a stable Style Guide (13 pages)**
- **Transcribed first 13 hrs of 20-hr benchmark**
  - **Some obvious disfluencies were missed**
- **Re-transcription of all 20 hrs due May 16**
  - **Will perform training with this data and test effect on recognition accuracy**

Sample Transcript

L: Oh, so you think it was fear that kept Iraq from using it.

R: Right. And what happens is --

L: But yet tha- --

R: What happens is suppose they get Saddam Hussein, which they eventually will, he's got one less thing to go against him. I mean if he were to use that, he might as well commit suicide because he's going to be captured and, you know. But the ~US also makes a lot of, uh, you know, treaties with other people. Like saying, "Okay, if you give up then you can come live in our country and we'll take care of you". Like Marcos, right?

L: Yeah.

R: "We'll overthrow you but, yeah, you can still come live here", you know.

L: Right. I don't think they've done that to Saddam Hussein yet.

R: [LAUGH] Ah, no. No.

Plan

- **Have WordWave transcribe 1700 hours from Fisher and Switchboard 2**
- **First 1000 hrs will be transcribed at a rate of ~100 hrs per week**
- **BBN will**
  - **Monitor the whole transcription process**
  - **Post-process data**
  - **Align transcripts to speech**
  - **Send to LDC for distribution**